BI 27048 (5)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 July 2002 (04.07.2002)

PCT

(10) International Publication Number
WO 02/052546 A1

(51) International Patent Classification[7]: G10L 21/02

(21) International Application Number: PCT/CN00/00733

(22) International Filing Date:
27 December 2000 (27.12.2000)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **PI, Xiaobo** [CN/CN]; Room 205, Building 901, Jin Song Road, Beijing 100021 (CN). **JIA, Ying** [CN/CN]; Room 415, Building 56, Bei Li Shi Road, Beijing (CN).

(74) Agent: **CCPIT PATENT AND TRADEMARK LAW OFFICE**; 8th Floor, 2 Fuchengmenwai St., Beijing 100037 (CN).

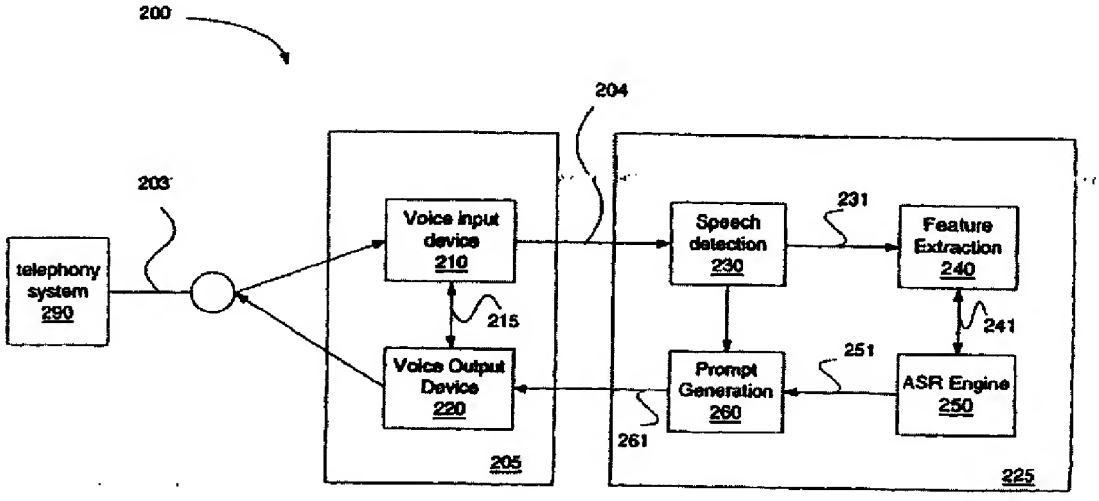
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: VOICE BARGE-IN IN TELEPHONY SPEECH RECOGNITION



(57) Abstract: An interactive voice response system is described that supports full duplex data transfer to enable the playing of a voice prompt to a user of telephony system while the system listens for voice barge-in from the user. The system includes a speech detection module that may utilize various criteria such as frame energy magnitude and duration thresholds to detect speech. The system also includes an automatic speech recognition engine. When the automatic speech recognition engine recognizes a segment of speech, a feature extraction module may be used to subtract a prompt echo spectrum, which corresponds to the currently playing voice prompt, from an echo-dirtied speech spectrum recorded by the system. In order to improve spectrum subtraction, an estimation of the time delay between the echo-dirtied speech and the prompt echo may also be performed.

WO 02/052546 A1

VOICE BARGE-IN IN TELEPHONY SPEECH RECOGNITION

FIELD OF THE INVENTION

The present invention relates to the field of speech recognition and, in particular, to voice barge-in for speech recognition based telephony applications.

BACKGROUND OF THE INVENTION

Speech recognition based telephony systems are used by businesses to answer phone calls with a system that engages users in natural language dialog. These systems use interactive voice response (IVR) telephony applications for a spoken language interface with a telephony system. IVR applications enable users to interrupt the system output at any time, for example, if the output is based on an erroneous understanding of a user's input or if it contains superfluous information that a user does not want to hear. Barge-in allows a user to interrupt a prompt being played using voice input. Enabling barge-in may significantly enhance the user's experience by allowing the user to interrupt the system prompt, whenever desired, in order to save time. Without barge-in, a user may react only when the system prompt completes, otherwise the user's input is ignored by the system. This may be very inconvenient to the user, particularly when the prompt is long and the user already knows the prompt message.

In today's touch tone based IVR systems, barge-in is widely adopted. However, for speech recognition based IVR systems, barge-in poses to be a much greater challenge due to background noise and echo from a prompt that may be transmitted to a voice recognition system.

One method of barge-in, referred to as key barge-in, is to stop playing a prompt and be ready to process a user's speech after the user presses a special key, such as the "#" or "*" key. One problem with such a method is that the user must be informed of how to use it. As such, another prompt may need to be added to the system, thereby undesirably increasing the amount of user interaction time with the system.

Another method of barge-in, referred to as voice barge-in, enables a user to speak directly to the system to interrupt the prompt. Figure 1 illustrates how barge-in occurs during prompt play in a voice barge-in system. Such a method uses speech detection to detect a user's speech while the prompt is playing. Once the user' speech is detected in the incoming data, the system stops playing and immediately begins a record phase in which the incoming data is made available to a speech recognition engine. The speech recognition engine processes the user's speech.

Although, such a method may provide a better solution than key barge-in, the voice barge-in function of current IVR systems has several problems. One problem with current IVR systems is that the computer-telephone cards used in these systems may not support full-duplex data transfer. Another problem with current IVR systems is that they may not be able to detect speech robustly from background noise, non-speech sounds, irrelevant speech and/or prompt echo. For example, the prompt echo that resides in these systems may significantly degrade speech quality. Using traditional adaptive filtering methods to remove near-end prompt echo may significantly degrade the performance of automatic speech recognition engines used in these systems.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

Figure 1 illustrates barge-in during prompt play in a voice barge-in system.

Figure 2 illustrates one embodiment of an interactive voice response telephony system.

Figure 3 illustrates one embodiment of a method of implementing an interactive voice response system.

Figure 4 illustrates one embodiment of speech detection in an input signal.

Figure 5 illustrates one embodiment of a feature extraction method.

Figure 6 illustrates an embodiment of a feature extraction method for a particular feature.

DETAILED DESCRIPTION

In the following description, numerous specific details are set forth such as examples of specific systems, components, modules, etc. in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that these specific details need not be employed to practice the present invention. In other instances, well known components or methods have not been described in detail in order to avoid unnecessarily obscuring the present invention.

The present invention includes various steps, which will be described below. The steps of the present invention may be performed by hardware components or may be embodied in machine-executable instructions, which may be used to cause a general-purpose or special-purpose processor programmed with the instructions to perform the steps. Alternatively, the steps may be performed by a combination of hardware and software.

The present invention may be provided as a computer program product, or software, that may include a machine-readable medium having stored thereon instructions, which may be used to program a computer system (or other electronic devices) to perform a process according to the present invention. A machine readable medium includes any mechanism for storing or transmitting information in a form (e.g., software) readable by a machine (e.g., a computer). The machine-readable medium may includes, but is not limited to, magnetic storage medium (e.g., floppy diskette); optical storage medium (e.g., CD-ROM); magneto-optical storage medium; read only memory (ROM); random access memory (RAM); erasable programmable memory (e.g., EPROM and EEPROM); flash memory; electrical, optical, acoustical or other form of propagated signal (e.g., carrier waves, infrared signals, digital signals, etc.); or other type of medium suitable for storing electronic instructions.

Figure 2 illustrates one embodiment of an interactive voice response telephony system. IVR system 200 allows for a spoken language interface with

telephony system 290. IVR system 200 supports voice barge-in by enabling a user to interrupt a prompt being played using voice input. In one embodiment, IVR system 200 includes interface module 205 and voice processing module 225. Interface module 205 provides interface circuitry for direct connection of voice processing module 225 with line 203 carrying voice data. Line 203 may be an analog or a digital line.

Interface module 205 includes voice input device 210 and voice output device 220. Voice input device 210 and voice output device 220 may be routed together using bus 215 to support full-duplex data transfer. Voice input device 210 provides for voice data transfer from telephony system 290 to voice processing module 225. Voice output device 220 provides for voice data transfer from voice processing module 225 to telephony system 290. For example, voice output device 220 may be used to play a voice prompt to a user of telephony system 290 while voice input device 210 is used to listen for barge-in (e.g., voice or key) from a user.

In one embodiment, for example, voice devices 210 and 220 may be Dialogic D41E cards, available from Dialogic Corporation of Parsippany, NJ. Dialogic's SCbus routing function may be used to establish communications between the Dialogic D41E cards. In alternative embodiment, voice devices from other manufacturers may be used, for example, cards available from Natural Microsystems of Framingham, MA.

In one embodiment, voice processing module 225 may be implemented as a software processing module. Voice processing module 225 includes speech detection module 230, feature extraction module 240, automatic speech recognition (ASR) engine 250, and prompt generation module 260. Speech detection module 230 may be used to detect voice initiation in the data signal received from voice input device 210. Feature extraction module 240 may be used to extract features used by ASR engine 250 and remove prompt from input signal 204. A feature is a representation of a speech signal that is suitable for automatic speech recognition. For example, a feature may be Mel-Frequency Cepstrum Coefficients (MFCC) and their first and second order derivatives, as

discussed below in relation to Figure 6. As such, feature extraction may be used to obtain a speech feature from the original speech signal waveform.

ASR engine 250 provides the function of speech recognition. Input 231 to ASR engine 250 contains vectors of speech. ASR engine 250 outputs 241 a recognition result as a word string. When ASR engine 250 recognizes a segment of speech, according to a particular prompt that is playing, feature extraction module 240 cleans up the speech containing data signal. For example, feature extraction module 240 may subtract the corresponding prompt echo's spectrum from the echo-dirtied speech spectrum. In one embodiment, ASR engine may be, for example, an Intel Speech Development Toolkit (ISDT) engine available from Intel Corporation of Santa Clara, CA. In alternative embodiment, another ASR engine may be used, for example, ViaVoice available from IBM of Armonk, N.Y. ASR engines are known in the art; accordingly, a detailed discussion is not provided.

Prompt generation module 260 generates prompts using a text-to-speech (TTS) engine that converts text input into speech output. For example, the input 251 to prompt generation module 260 may be a sentence text and the output 261 is a speech waveform of the sentence text. TTS engines are available from industry manufacturers such as Lucent of Murray Hill, NJ and Lernout & Hauspie of Belgium. In an alternative embodiment, a custom TTS engine may be used. TTS engines are known in the art; accordingly, a detailed discussion is not provided.

After prompt waveform is generated, prompt generation module 260 plays a prompt through voice output device 220 to the user of telephony system 290. It should be noted that in an alternative embodiment, the operation of voice processing module 225 may be implemented in hardware, for example, is a digital signal processor.

Referring again to speech detection module 230, in one embodiment, two criteria may be used to determine if input signal 204 contains speech. One criteria may be based on frame energy. A frame is a segment of input signal 204. Frame energy is the signal energy within the segment. In one embodiment, if a

segment of the detected input signal 204 contains speech, then it may be assumed that a certain number of frames of a running window of frames will have their energy levels above a predetermined minimum energy threshold. The window of frames may be either sequential or non-sequential. The energy threshold may be set to account for energy from non-desired speech, such as energy from prompt echo.

In one embodiment, for example, a frame may be set to be 20 milliseconds (ms), where speech is assumed to be short-time stationary up to 20ms; the number of frames may be set to be 8 frames; and the running window may be set to be 10 frames. If, in this running window, the energy of 8 frames is over the predetermined minimum energy threshold then the current time may be considered as the start point of the speech. The energy threshold may be based on, for example, an average energy of prompt echo that is the echo of prompt currently being played. In this manner, the frame energy threshold may be set dynamically. According to different echos of prompt, the frame energy threshold may be set as the average energy of the echo. The average energy of prompt echo may be pre-computed and stored when a prompt is added into system 200.

Another criteria that may be used to determine if input signal 204 contains speech is the duration of input signal 204. If the duration of input signal 204 is greater than a predetermined value then it may be assumed that input signal 204 contains speech. For example, in one embodiment, it is assumed that any speech event lasts at least 300 ms. As such, the duration value may be set to be 300 ms.

After a possible start point of speech is detected, speech detection module 230 attempts to detect the end point of the speech using the same method as detecting the start point. The start point and the end point of speech are used to calculate the duration. Continuing the example, if the speech duration is over 300 ms then the possible start point of speech is a real speech start point and the current speech frames and successive speech frames may be sent to feature extraction module 240. Otherwise, the possible start point of speech is not a real

ID: <WO 02052546A1 | -

start point of speech and speech detection is reset. This procedure lasts until an end point of speech is detected or input signal 204 is over a maximum possible length.

Speech detection module 230 may also be used to estimate the time delay of the prompt echo in input signal 204 if an echo cancellation function of system 200 is desired. While a prompt is added in system 200, its waveform may be generated by prompt generation module 260. The waveform of the prompt is played once so that its echo is recorded and stored. When processing an input signal, correlation coefficients between input signal 204 and the stored prompt echo is calculated with the following equation:

$$C(\tau) = \sum_{t=1}^{T} S(t+\tau) \times E(t)$$

where C is the correlation coefficients; S is input signal 204, E is the prompt echo, T is the echo length, and τ is the time delay estimation of echo. The value of τ may range from zero to the maximum delay time (e.g., 200 ms). After C is computed, the maximum value of C in all τ is found. This value of τ is the time-delay estimation of echo. This value is used in the feature extraction module 240 when performing spectrum subtraction of the prompt echo spectrum to remove prompt echo from the input signal 204 having echo dirtied speech, as discussed below.

Figure 3 illustrates one embodiment of a method of implementing an interactive voice response system. A prompt echo waveform 302 and an input signal 301 are received by the system. In one embodiment, a speech detection module may estimate the time delay time of a feature in input signal 301, step 310. The speech detection module may also be used to detect the existence of speech in input signal 302, in step 320. The existence of speech may be based on various criteria, such as amount of frame energy of the input signal and the duration of frame energy, as discussed below in relation to Figure 4.

In step 330, feature extraction may be used to obtain a speech feature from the original speech signal waveform. In one embodiment, prompt echo may be removed from input signal 301, using spectrum subtraction, to facilitate the

recognition of speech in the input signal. After feature extraction is performed, speech recognition may be performed on input signal 301, step 340. A prompt may then be generated, step 350, based on the recognized speech.

Figure 4 illustrates one embodiment of speech detection in an input signal. In one embodiment, the frame energy of an input signal may be used to determine if the input signal contains speech. An assumption may be made that if the energy of the input signal, over a certain period of time, is above a certain threshold level, then the signal may contain speech.

Thus, in one embodiment, an energy threshold for the input signal may be set, step 410. The energy threshold is set higher than the prompt echo energy so that the system will not consider the energy of prompt echo in the input signal to be speech. In one embodiment, the energy threshold may be based on an average energy of the prompt echo that is the echo of the prompt currently playing during the speech detection. The energy of the input signal is measured over a predetermined time period, step 420, and compared against the energy threshold.

The input signal may be measured over time segments, or frames. In one embodiment, for example, a frame length of an input signal may be 20 milliseconds in duration where speech is assumed to be a short-time stationary event up to 20 milliseconds. In step 430, the number of energy frames containing energy above the threshold is counted. If the energy of the input signal over a predetermined number of frames (e.g., 8 frames) is greater than the predetermined energy threshold, then the input signal may be considered to contain speech with that point of time as the start of speech, step 440.

In one embodiment, the energy of the input signal may be monitored over a running window of time. If in this running window (e.g., 10 frames) there is the predetermined number of frames (e.g., 8 frames) over the predetermined energy threshold, then that point of time may be considered as the start of speech.

In an alternative embodiment, another method of detecting the start of speech may be used. For example, the rate of input signal energy crossing over

the predetermined threshold may be calculated. If the measure rate exceeds a predetermine rate, such as a zero-cross threshold rate, then the existence and start time of speech in the input signal may be determined.

If no speech is detected in the input signal, then a determination may be made whether the period of silence (i.e., non-speech) is too long, step 445. If a predetermined silence period is not exceeded, then the system continues to monitor the input signal for speech. If the predetermined silence period is exceeded, then the system may end its listening and take other actions, for example, error processing (e.g., close the current call), step 447.

In one embodiment, the duration of frame energy of an input signal may also be used to determine if the input signal contains speech. A possible start point of speech is detected as described above in relation to steps 410 through 440. After a possible start point of speech is detected, then the end point of the speech is detected to determine the duration of speech, step 450. In one embodiment, the end point of speech may be determined in a manner similar to that of detecting the possible start point of speech. For example, the energy of the input signal may be measured over another predetermined time period and compared against the energy threshold. If the energy over the predetermined time period is less than the energy threshold then the speech in the input signal may be considered to have ended. In one embodiment, the predetermined time in the speech end point determination may be the same as the predetermined time in the speech start point determination. In an alternative embodiment, the predetermined time in the speech end point determination may be different than the predetermined time in the speech start point determination.

Once the end point of speech is determined, the duration of the speech is calculated, step 460. If the duration is above a predetermined duration threshold, then the possible start point of speech is a real speech start point step 470. In one embodiment, for example, the predetermined duration threshold may be set to 300 ms where it is assumed that any anticipated speech event lasts for at least 300 ms.

Otherwise, the possible start point of speech is not a real start point of speech and the speech detection may be reset. This procedure lasts until an end point of speech is detected or the input signal is over a maximum possible length, step 480.

Figure 5 illustrates one embodiment of a feature extraction method. In one embodiment, an input signal and a prompt echo waveform are received, steps 515 and 525, respectively. A Fourier transformation is performed to obtain a speech spectrum from the input signal, step 510. A Fourier transformation may also be performed on the echo waveform to generate a prompt echo spectrum, step 526.

In one embodiment, the prompt echo spectrum is shifted according to a time delay estimated between the input signal and the prompt echo waveform, step 519. The prompt echo spectrum is computed and subtracted from the speech spectrum, step 520. Afterwards, the Cepstrum coefficients may be obtained for use by ASR engine 250 of Figure 2 in performing speech recognition, step 530.

In one embodiment, feature extraction involves the cancellation of echo prompt from the input signal, as discussed below in relation to Figure 6. When ASR engine 250 of Figure 2 recognizes a segment of speech, feature extraction may be used to subtract a prompt echo spectrum that corresponds to the currently playing prompt from echo-dirtied speech spectrum. In order to improve spectrum subtraction, an estimation of the time delay between the echo-dirtied speech and the recorded echo may be performed by speech detection module 230 of Figure 2.

Figure 6 illustrates an embodiment of a feature extraction method for a particular feature. In one embodiment, Mel-Frequency Cepstrum Coefficients (MFCC) may be used to in performing speech recognition. Using a MFCC generation procedure, a Hamming window is added to the frame segment set for speech (e.g., 20 ms), step 610. A Fast Fourier Transform (FFT) is calculated to obtain the speech spectrum, step 620. If the echo spectrum subtraction function is enabled, shift the echo waveform according to the time delay then compute

the echo spectrum and subtract the echo spectrum from the input signal spectrum, step 630. Next perform a logarithmic operation on the speech spectrum, step 640. Perform Mel-scale warping to reflect the non-linear perceptual characteristics of human hearing, step 650. Perform Inverse Discrete Time Transformation (IDCT) to obtain the Cepstrum coefficients, step 660. The resulting feature is a multiple (e.g., 12 dimension) vector. These parameters form the base feature of MFCC.

In one embodiment, the first and second derivatives of the base feature are added to be the additional dimensions (the 13th to 24th and 25th to 36th dimensions, respectively), to account for a change of speech over time. By using near-end prompt echo cancellation, the performance of the ASR engine 250 of Figure 2 may be improved. In one embodiment, for example, the performance of the ASR engine 250 of Figure 2 may improve by greater than 6%.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the following claims. The specification and drawings are, accordingly, to be regarded in an illustrative sense rather than a restrictive sense.

CLAIMS

What is claimed is:

1. A method, comprising:

detecting an existence of speech in an input signal; and

removing a prompt echo from the input signal using spectrum subtraction.

2. The method of claim 1, wherein removing the prompt echo comprises extracting a feature from input signal to generate a plurality of coefficients and wherein the method further comprises:

performing speech recognition on the input signal using the plurality of coefficients; and

generating a prompt in response to particular speech recognized in the input signal.

3. The method of claim 1, wherein the existence of speech is detected based on a predetermined energy in a plurality of segments of the input signal.

4. The method of claim 3, wherein the existence of speech is detected based on a predetermined duration of the plurality of segments having the predetermined energy.

5. The method of claim 1, wherein detecting the existence of speech in the input signal comprises:

setting an energy threshold for the input signal, the input signal having a plurality of segments; and

determining a start point of speech in the input signal, comprising:

measuring an energy of the input signal for a first predetermined time; and

determining whether the energy of the input signal for the first predetermined time is greater than the energy threshold.

6. The method of claim 5, wherein detecting the existence of speech in the input signal further comprises:

measuring a duration of the energy above the energy threshold; and

determining whether the duration is greater than a predetermined duration threshold.

7. The method of claim 1, wherein removing the prompt echo from the input signal comprises:

estimating a time delay between the input signal and the echo;

obtaining a speech spectrum from the input signal, the speech spectrum including the echo;

shifting the echo according to the time delay;

computing an echo spectrum using the shifted echo; and

subtracting the echo spectrum from the speech spectrum.

8. The method of claim 1, wherein removing the prompt echo from the input signal comprises:

estimating a time delay between the input signal and the prompt echo; and

removing the prompt echo from the input signal based on the time delay.

9. The method of claim 8, wherein removing prompt echo comprises:

generating a prompt echo spectrum;

calculating a Fast Fourier Transform using the input signal to obtain a speech spectrum;

subtracting the prompt echo spectrum from the speech spectrum using the estimated time delay.

10. The method of claim 9, further comprising:

performing an inverse DCT to obtain Cepstrum coefficients;

performing a logarithm on the speech spectrum; and

performing Mel-scale warping on the logarithm of the speech spectrum.

11. A method, comprising:

setting an energy threshold for a signal having a plurality of segments; and

determining a start point of speech in the signal, comprising:

measuring an energy of the signal for a first predetermined time; and

determining whether the energy of the signal for the first predetermined time is greater than the energy threshold.

12. The method of claim 11, wherein setting an energy threshold comprises:

measuring a prompt echo energy; and

setting the energy threshold above the prompt echo energy.

13. The method of claim 11, further comprising determining an end point of the speech.

14. The method of claim 13, wherein determining the end point of the speech comprises:

measuring the energy of the signal for a second predetermined time; and

determining whether the energy of the signal for the second predetermined time is less than the energy threshold.

15. The method of claim 14, further comprising:

calculating a duration based on the start point and the end point; and

determining whether the duration is greater than a predetermined duration threshold.

16. The method of claim 14, wherein the first and second predetermined times are the same.

17. The method of claim 13, wherein determining the end point of the speech comprises:

measuring the energy of the signal for a second predetermined time;

calculating a rate of energy crossing over the energy threshold; and

determining whether the rate is greater than a predetermined rate.

18. The method of claim 17, further comprising:

calculating a duration based on the start point and the end point; and

determining whether the duration is greater than a predetermined duration threshold.

19. The method of claim 11, wherein the determining whether the energy of the signal for the first predetermined time is greater than the energy threshold is performed over a running window of time.

20. The method of claim 16, wherein the first predetermined number of segments is eight and the running window of segments is ten.

21. The method of claim 20, wherein a segment of the plurality of segments is 20 milliseconds.

22. A machine readable medium having stored thereon instructions, which when executed by a processor, cause the processor to perform the following, comprising:

detecting the existence of speech in an input signal; and

removing a prompt echo from the input signal using spectrum subtraction.

23. The machine readable medium of claim 22, wherein removing the prompt echo from the input signal causes the processor to perform the following, comprising:

estimating a time delay between the input signal and the prompt echo;

obtaining a speech spectrum from the input signal, the speech spectrum including the prompt echo;

shifting the prompt echo according to the time delay;

computing a prompt echo spectrum using the shifted echo; and

subtracting the prompt echo spectrum from the speech spectrum.

24. The machine readable medium of claim 22, wherein the existence of speech is detected based on a predetermined energy in a plurality of segments of the input signal and a predetermined duration of the plurality of segments having the predetermined energy.

25. A machine readable medium having stored thereon instructions, which when executed by a processor, cause the processor to perform the following, comprising:

setting an energy threshold for a signal having a plurality of segments; and

determining a start point of speech in the signal, comprising:

measuring an energy of the signal for a first predetermined time; and

determining whether the energy of the signal for the first predetermined time is greater than the energy threshold.

26. The machine readable medium of claim 25, wherein the processor further performs the following, comprising:

determining an end point of speech in the signal, comprising:

measuring the energy of the signal for a second predetermined time; and

determining whether the energy of the signal for the second predetermined time is less than the energy threshold.

27. The machine readable medium of claim 25, wherein the processor further performs the following, comprising:

calculating a duration based on the start point and the end point; and

determining whether the duration is greater than a predetermined duration threshold.

28. An apparatus, comprising:

a voice processing module; and

a voice interface device coupled to the voice processing module, the voice interface device comprising:

a voice input device; and

a voice output device coupled to the voice input device to support full-duplex data transfer between the voice interface device and a telephony system.

29. The apparatus of claim 28, wherein the voice processing module comprises a digital signal processor.

30. The apparatus of claim 28, wherein the voice processing module comprises processing software and wherein the processing software comprises:

a speech detection module;

a feature extraction module;

an automatic speech recognition engine; and

a prompt generation module.

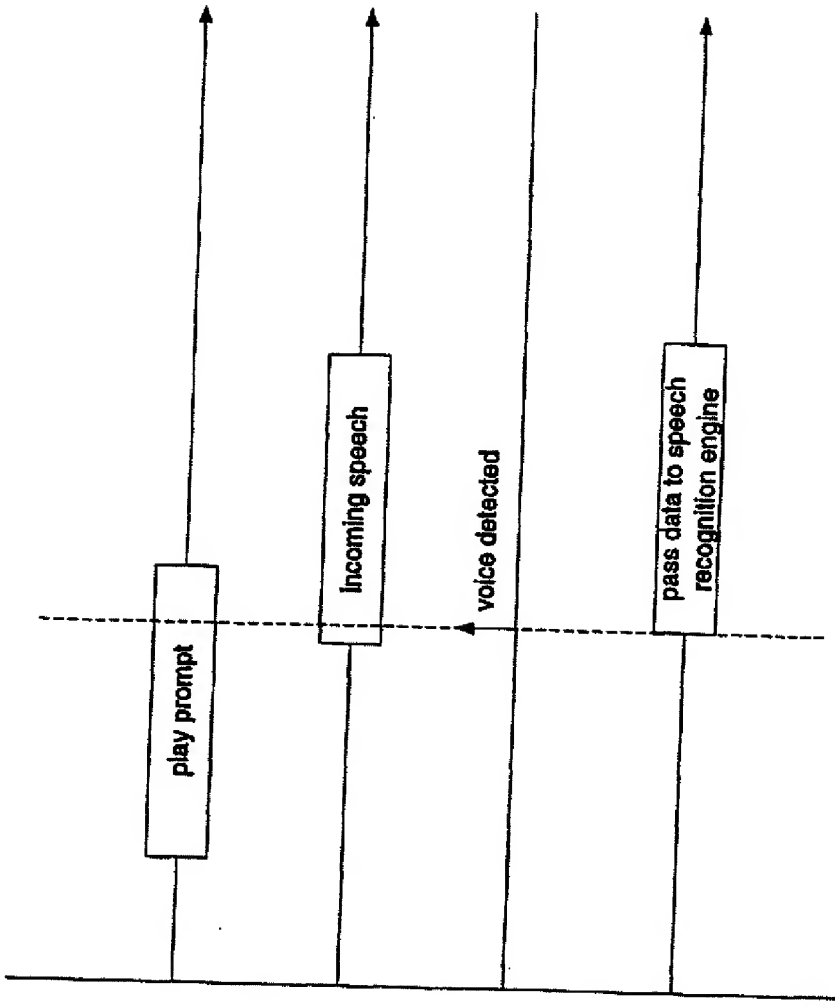

play prompt
incoming speech
voice detected
pass data to speech recognition engine

Figure 1

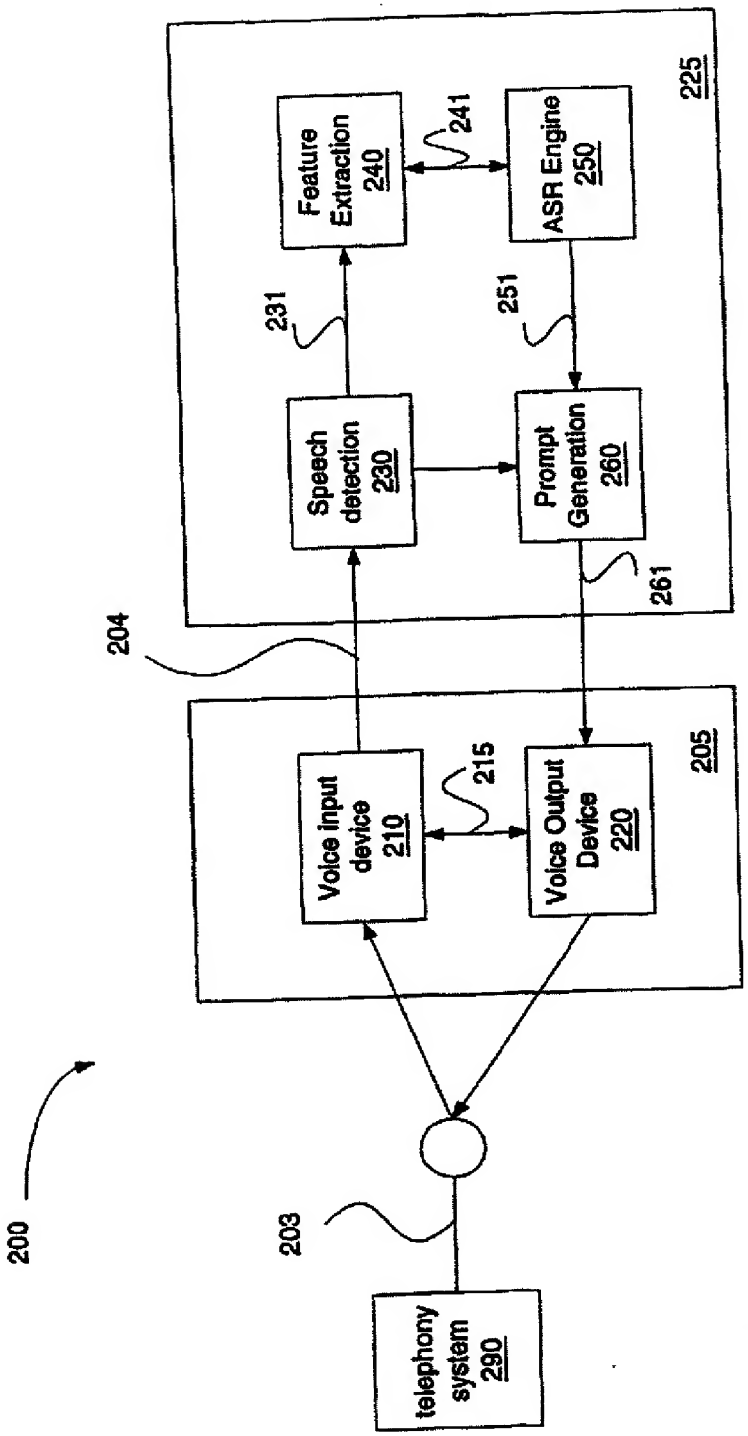

225
Feature Extraction 240
241
ASR Engine 250
231
251
Speech detection 230
Prompt Generation 260
261
204
205
Voice input device 210
215
Voice Output Device 220
200
203
telephony system 290

Figure 2

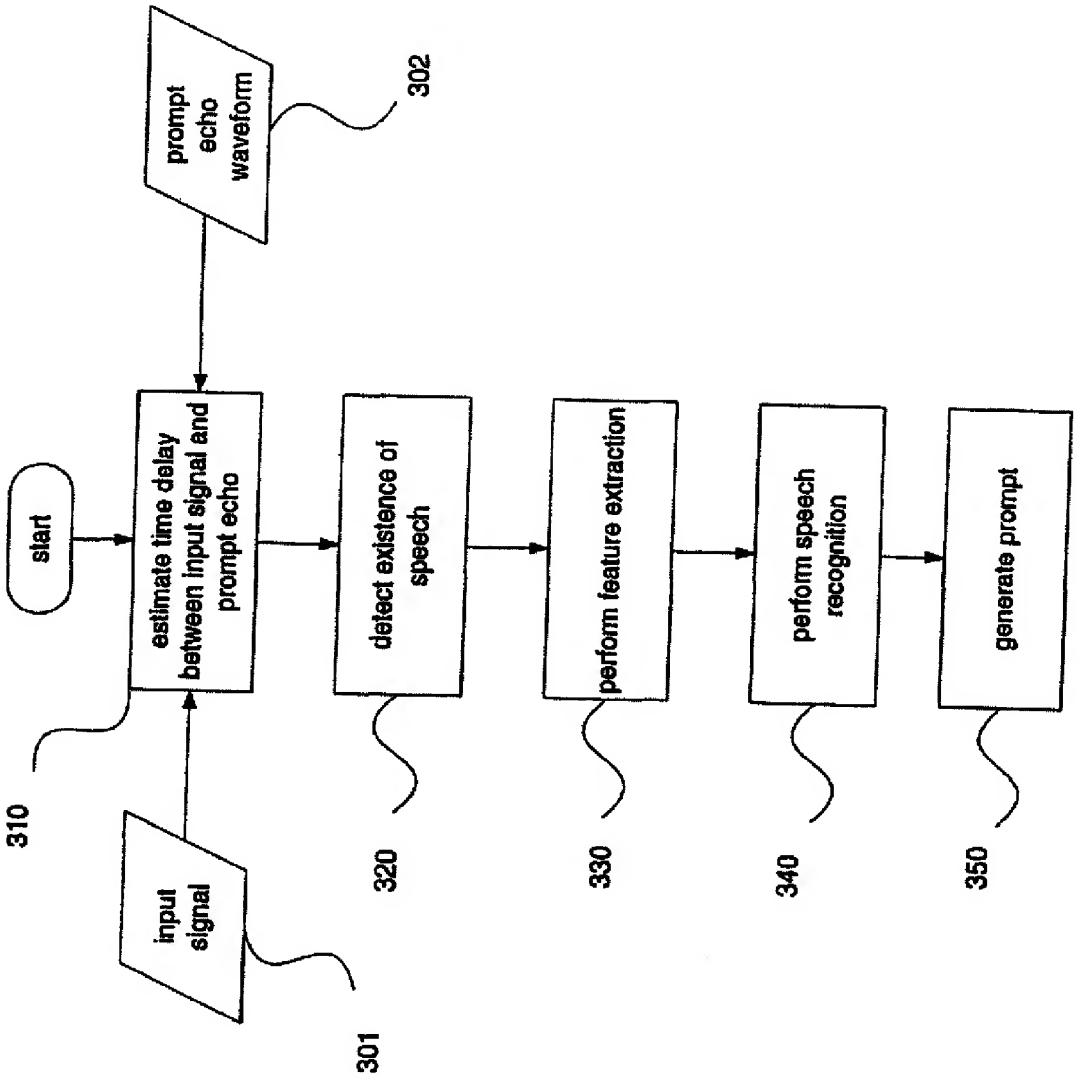

prompt echo waveform
302
start
estimate time delay between input signal and prompt echo
detect existence of speech
perform feature extraction
perform speech recognition
generate prompt
310
320
330
340
350
input signal
301

Figure 3

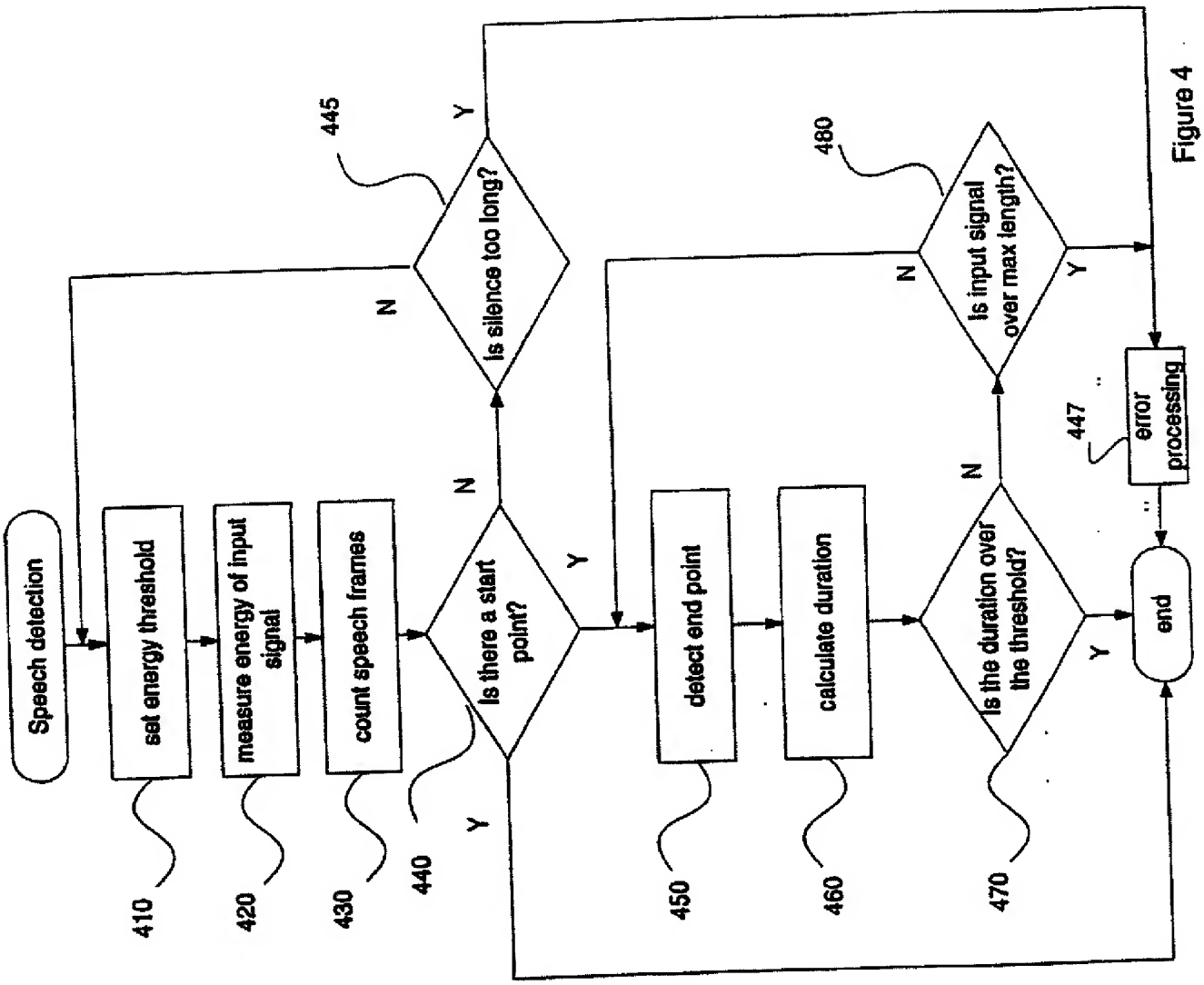

Speech detection
set energy threshold
measure energy of input signal
count speech frames
Is there a start point?
Is silence too long?
detect end point
calculate duration
Is the duration over the threshold?
Is input signal over max length?
error processing
end
410
420
430
440
445
447
450
460
470
480
Y
N

Figure 4

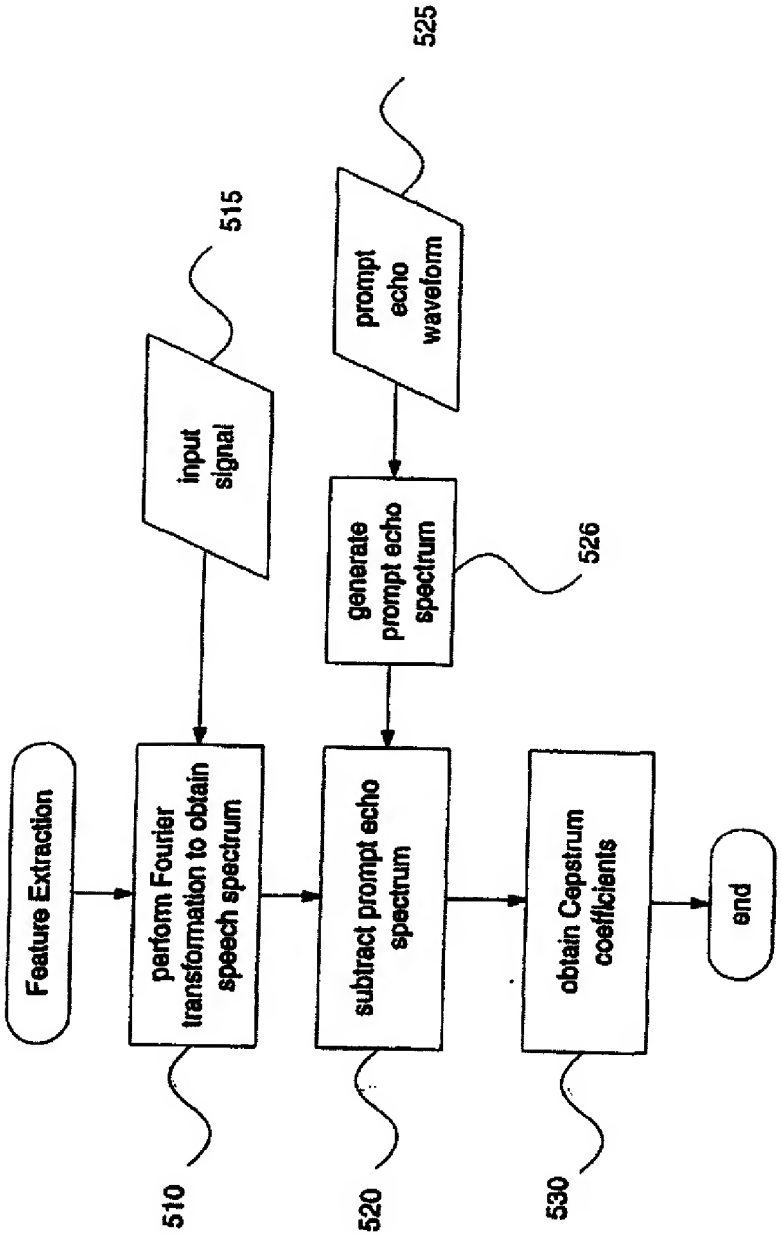

Feature Extraction
perform Fourier transformation to obtain speech spectrum
input signal
515
prompt echo waveform
525
generate prompt echo spectrum
526
subtract prompt echo spectrum
obtain Cepstrum coefficients
end
510
520
530

Figure 5

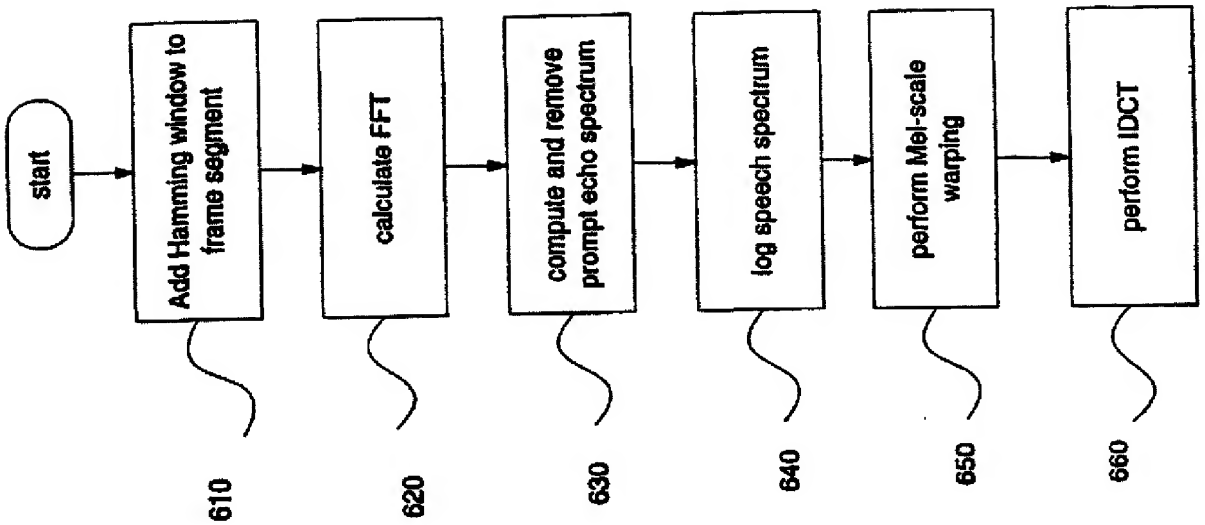

start
Add Hamming window to frame segment
calculate FFT
compute and remove prompt echo spectrum
log speech spectrum
perform Mel-scale warping
perform IDCT
610
620
630
640
650
660

Figure 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN00/00733

A. CLASSIFICATION OF SUBJECT MATTER

IPC7 G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7 G10L21/02,21/00,15/02,15/20,11/02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Chinese Patent Document (1985-)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI,EPODOC,PAJ,CNPAT

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US5991726(Peter Immarco et al)23.Nov 1999(23.11.99) Fig.3 and the description thereof	1,22,25,28
A	US5956675(Anand Rangaswamy Setlur)21.Sep 1999(21.09.99) Fig3,4 and the description thereof	1,22,25,28
A	US6061651(John N.Nguyen)09.May 2000(09.05.00) the whole document	1,22,25,28
A	US5765130(Jon N.Nguyen) 09.Jun 1998(09.01.98) the whole document	1,22,25,28
A	EP0736995(TEXAS INSTRUMENTS) 09.Oct 1996(09.10.96) Fig.2-4B and the description thereof	1,22,25,28

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
12.Jul 2001(12.07.01)

Date of mailing of the international search report
02 AUG 2001 (02.08.01)

Name and mailing address of the ISA/CN
6 Xitucheng Rd., Jimen Bridge, Haidian District,
100088 Beijing, China
Facsimile No. 86-10-62019451

Authorized officer

Telephone No. 86-10-62093916

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN00/00733

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US5991726	23.11.99	None	
US5956675	21.09.99	KR99014292	25.02.99
		EP0895224	03.02.99
		JP11095791	09.04.99
		CA2238642	31.01.99
US6061651	09.05.00	None	
US5765130	09.06.98	None	
EP0736995	09.10.96	JP9008708	10.01.97
		US5708704	13.01.98